# 6. EXPLORATORY METHODS DEVELOPMENT FOR ANALYSIS OF GENOMIC DATA FOR APPLICATION TO RISK ASSESSMENT

## 6.1. OBJECTIVES AND INTRODUCTION

The overall goal of this chapter is to describe exploratory methods developed for analyzing and applying toxicogenomic data in risk assessment. The three objectives of the methods development projects were to

1. *Explore the development of new methods to analyze microarray data for application to risk assessment.*

   The motivation was to develop methods for performing gene expression analyses of microarray data for use in risk assessment. Microarray studies for basic research purposes do not necessarily require as high a level of stringency as for risk assessment purposes because the analyses are often performed to generate hypotheses (e.g., for developing MOA hypotheses) that are subsequently tested in additional studies.

2. *Utilize existing DBP genomic data to develop a temporal gene network model for use in risk assessment.*

   We asked whether there are data to understand gene expression changes over time. By modeling the gene and pathway interactions across the critical window of exposure to DBP, it may be possible to understand the relationships among genes and pathways over time, and possibly, to identify the initiating event(s) for the decreases in fetal testicular T or *Insl3* expression. Identifying the initiating event would be very useful to risk assessment, as this would provide a biologically significant gene whose expression is critical to the outcome.

3. *Utilize genomic and other molecular data to address the Case Study Question: Do the toxicogenomic data inform interspecies differences in TD?*

   We utilized the available gene sequence data, protein sequence, and pathway cross-species data to assess the rat-to-human conservation of the genes involved in the steroidogenesis pathway that underlie the reduced fetal testicular T MOA for DBP.

The work to address the objectives of this chapter is the result of a collaborative effort among scientists at the STAR Bioinformatics Center at UMDNJ and Rutgers, and the EPA. The analyses were performed at Rutgers University.

The work presented in this chapter is highly technical and thus, is intended to be beneficial to scientists with expertise in bioinformatics. The technical details of the analyses are

provided in order that scientists could apply these methods to their work. Such an approach will allow the risk assessor proficient in microarray analysis methodology an opportunity to apply these methods. The last section of this chapter (Section 6.4) summarizes the findings for a scientific audience that does not have a strong background in microarray analysis methods.

## 6.2. PATHWAY ANALYSIS AND GENE INTERACTIONS AFTER *IN UTERO* DBP EXPOSURE

### 6.2.1. Pathway Activity Approach

Usually, to identify significant biological pathways from transcriptional data, pathway analysis is performed after determining the DEGs using a statistical filter. Two examples of this approach are described in Chapter 5 (Section 5.5). An alternative approach is the use of "pathway scoring" methods, which begin with projecting gene expression changes onto pathways (Rahnenfuhrer et al., 2004; Moothaet al., 2003; Hanisch et al., 2002). The main advantage of applying pathway scoring methods to microarray data is that changes can be identified at the pathway level that may not be detected by first identifying individual DEGs. Most of these methods calculate the average correlation between pairs of genes within pathways (Rahnenfuhrer et al., 2004; Sohler et al., 2004; Hanisch et al., 2002; Zien et al., 2000). Another pathway scoring method tests for association between gene expression and a phenotype (e.g., Gene Set Enrichment Analysis [GSEA]; Mootha et al., 2003). In GSEA, all genes are ranked with respect to some measure that quantifies the gene expression associated with a phenotype (i.e., differentiation between healthy vs. disease samples). Tomfohr et al. (2005) introduced a pathway-based approach that is similar in spirit to GSEA. Their method translates the overall gene expression levels within a pathway to a "pathway activity level," which is derived from singular value decomposition (SVD), described below. Hence, pathway activity levels can be used in the same kinds of applications as gene expression levels (Tomfohr et al., 2005). Tomfohr et al. (2005) compared their pathway activity method to GSEA using expression data from two different studies, one that studied Type 2 diabetes and one that studied the influence of cigarette smoke on gene expression in airway epithelia. They found similar results to those obtained using GSEA in the diabetes set, and further, improved results for identifying differentially expressed pathways in the cigarette smoke data.

We applied a pathway activity level approach to DBP microarray data. Since pathway activity levels are a reduced form of the overall gene expression matrix (represented by the

largest deviation in the overall gene expressions within a pathway) Alter et al. (2000) and Cangelosi (2007) raised the critical issue that pathway activity levels (represented by the largest deviation in the overall gene expressions within a pathway) may be attributed to random deviations in the data. Therefore, we use a significance analysis to distinguish the information captured by pathway activity levels from random deviation.

#### 6.2.1.1. *Significance Analysis of Pathway Activity Levels*

The procedure begins with mapping genes to the KEGG pathway database. The entire gene set represented by the Liu et al. (2005) data set (i.e., using the Affymetrix RAE230 A and B chips) maps to 199 pathways in the KEGG database with 4,772 associated genes.

Pathway activity formulation starts with SVD of the gene expression matrix of a given pathway. SVD involves a mathematical procedure that transforms a number of possibly correlated variables into a smaller number of uncorrelated variables. It mathematically transforms the data to a new coordinate system such that the greatest variance by any projection of the data lies on the first coordinate (called the eigenvector), the second greatest variance on the second coordinate, and so on. Associated with each of these coordinate eigenvectors is a weight term (called the eigenvalue) that represents the variance in the data. The eigenvalues are normalized such that they express the fraction of the variance along their corresponding eigenvector. In this study, SVD is used to calculate pathway activity levels for each experimental condition where each pathway activity level represents the most significant gene expression pattern within each pathway. The details of SVD analysis are as follows:
Using Eq. 6-1, let $\Xi_p(k,t)$ be the gene expression data associated with a given pathway, $p$, composed of $k$ genes measured at $t$ different conditions (time, treatment, dose, etc.), normalized (i.e., to a mean of zero mean and unit standard deviation). The SVD of $\Xi_p(k,t)$ is given as follows:

$$\Xi_p(k,t) = U_p(k,k) \times S_p(k,t) \times V_p(t,t)^I \tag{6-1}$$

Eq. 6-1states that the columns of the matrix $U_p(k,k)$ are the orthonormal eigenvectors of $\Xi_p(k,t)$. $S_p(k,t)$ is a diagonal matrix containing the associated eigenvalues, and the columns of the matrix $V_p(t,t)$ are projections of the associated eigenvectors of $\Xi_p(k,t)$. As the elements of $S_p(k,t)$ are

sorted from the highest to the lowest, the first row of $V_p(t,t)$ represents the most significant pattern within a pathway across different samples. Hence, $PAL_p$ is mathematically defined as the first vector of the $V_p(t,t)$ (given in Eq. 6-2 ).

$$PAL_p = V_p(n,1)^I \tag{6-2}$$

The fraction of the overall gene expression that is captured by $PAL_p$ is evaluated through Eq. 6-3.

$$f_P = \frac{S_p(1,1)^2}{\sum_{g=1}^{L} S_p(L,L)^2} \tag{6-3}$$

An additional analysis is needed to evaluate whether $PAL_p$ represents significant information about the pathway. As a standard procedure for evaluating significance of microarray data, random sampling is used. For each pathway, an equal number of gene expression values are permutated 1,000 times. The $p$-value is computed as the permutated $f_p$ that exceeded the actual $f_p$ ($p$-value < 0.05). Next, the pathways are filtered based on the associated $p$-value of their $f_p$ value.

We illustrate the importance of the significance analysis of $PAL_p$ in Figure 6-1 using the gene expression matrix for the tryptophan metabolism pathway. Panel A of Figure 6-1 depicts both the fraction of the overall gene expression captured by each eigenvector, $f_p$ , and the average fraction of the overall gene expression captured by each eigenvector of the randomized data. We observe that the $f_p$ value captured by the $PAL_p$ of the tryptophan metabolism pathway can be retrieved with a randomly selected gene set and thus, the tryptophan metabolism pathway is not significantly affected by DBP exposure. We applied a significance analysis of $PAL_p$ to improve the confidence of Tomfohr's pathway activity level formulation for further calculations.

#### 6.2.1.2. *Pathway Activity Analysis*

The main goal of pathway analysis is to identify significantly affected pathways, based on gene expression data, due to DBP exposure. For this purpose, as described above,

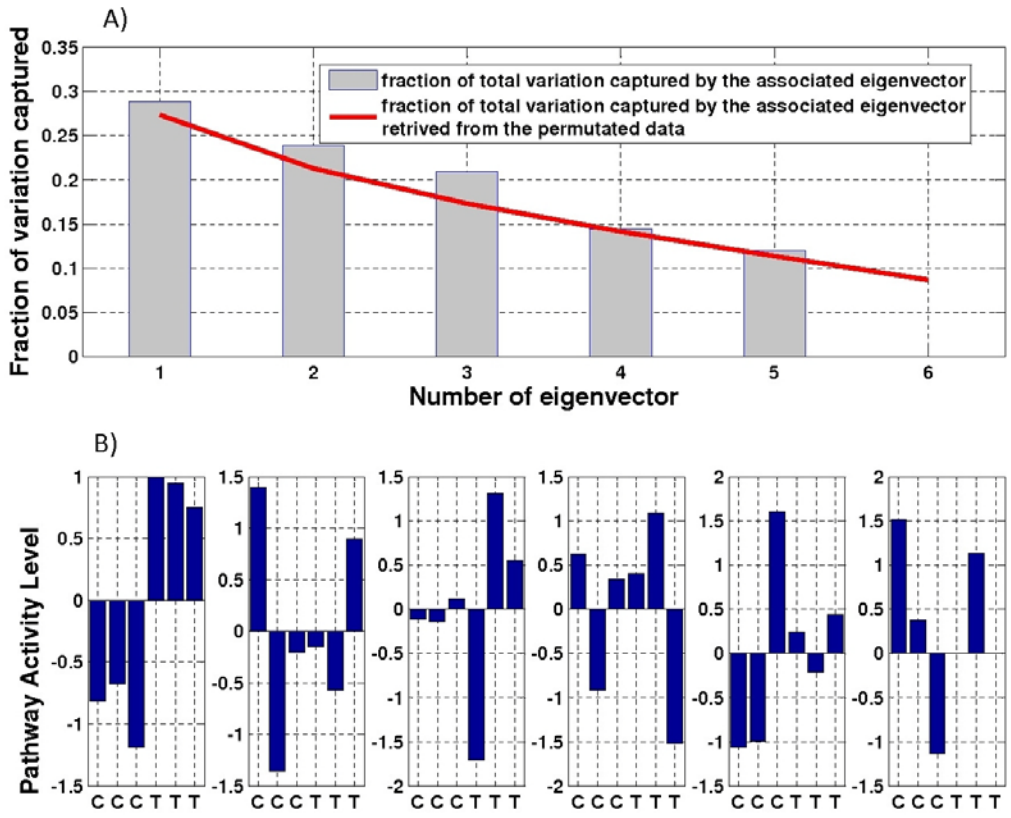


**Figure 6-1. An illustration of the adapted version of pathway activity level analysis for the tryptophan metabolism pathway, a nonactive pathway for DBP.** In panel A, the boxes indicate the variability in the actual gene expression data, associated with the tryptophan metabolism for each individual eigenvector. For comparison, the solid line represents the fraction of data variability captured by the corresponding eigenvectors when randomly generated data were used. No apparent distinction between the actual data and randomly generated data was identified, as quantified by the calculated *p*-value of 0.25. In panel B, the projection of the gene expression on each eigenvector is depicted for each sample of the control (C) and DBP-treated (T) groups. $PAL_p$ is the first vector that corresponds to the largest variation in the data.

overall gene expressions within a pathway are reduced to $PAL_p$. The differentiation between $PAL_p$ of different samples is denoted as pathway activity and is determined through a process analogous to SNR analysis.

If $n_1$ samples are associated with vehicle treatment (control) and $n_2$ samples with chemical treatment (DBP), then the activity levels associated with treatment groups are given in Eqs. 6-4 and 6-5, respectively.

$$PAL_p^1 = V_p(n_1, 1)^I \tag{6-4}$$

$$PAL_p^2 = V_p(n_2, 1)^I \tag{6-5}$$

Pathway activity is calculated using Eq. 6-6 where $\mu$ and $\sigma$ represent the mean and standard deviation respectively.

$$PA_p = \frac{\left|\mu\left(PAL_p^1\right) - \mu\left(PAL_p^2\right)\right|}{\sigma\left(PAL_p^1\right) + \sigma\left(PAL_p^2\right)} \tag{6-6}$$

A high pathway activity represents a better differentiation between control and treated pathway activity levels. The statistical significance of pathway activity is determined using the randomization process. For each pathway, an equal number of genes within a given pathway are randomly assigned and gene expression changes are generated (from the chip) 10,000 times. The *p*-value of the pathway activity is computed as the fraction of the randomized pathway activity that exceeded the actual pathway activity. In this analysis, the pathways that have both statistically significant (*p*-value < 0.05) pathway activity and pathway activity level are defined as "active" pathways.

"Active" pathways are those for which the overall change in gene expression in a pathway of treated samples compared to control samples was statistically significant. For example, an active pathway could be one for which gene expression was downregulated or turned off after DBP exposure. Alternatively, a pathway that is not identified as active would still have gene expression occurring, but might not exhibit a significant difference in gene expression following DBP exposure compared to the control samples. Thus, the term active does not refer to gene expression from a particular pathway. The algorithm for selecting active pathways using the pathway activity method is shown in Appendix B, Figure B-1.

We identified 15 active pathways from querying the KEGG database (see Table 6-1). The pathway activity method identified pathways such as biosynthesis of steroids (C21 Steroid hormone metabolism pathways known to be biologically relevant to T levels) as well as other pathways including butanoate metabolism, pyruvate metabolism, and biosynthesis of unsaturated fatty acids (PPAR signaling pathway and fatty acid metabolism).

**Table 6-1. The KEGG pathways ordered based on their *p*-value for pathway activity[a]**

| Pathway name | p-value of PA[b] | p-value of PAL[c] |
|---|---|---|
| Reductive carboxylate cycle (CO2 fixation) | <0.001 | 0.002 |
| Valine, leucine and isoleucine degradation | <0.001 | <0.001 |
| Biosynthesis of steroids | 0.001 | <0.001 |
| Citrate cycle (TCA cycle) | 0.002 | <0.001 |
| Glutathione metabolism | 0.002 | 0.006 |
| Tryptophan metabolism† | 0.002 | 0.250 |
| Pentose phosphate pathway | 0.002 | <0.001 |
| Glycolysis / Gluconeogenesis | 0.003 | <0.001 |
| Butanoate metabolism | 0.004 | 0.006 |
| Pyruvate metabolism | 0.004 | <0.001 |
| C21Steroid hormone metabolism | 0.006 | 0.048 |
| Glyoxylate and dicarboxylate metabolism† | 0.012 | 0.480 |
| Biosynthesis of unsaturated fatty acids | 0.012 | 0.048 |
| Fatty acid metabolism | 0.020 | 0.030 |
| Nicotinate and nicotinamide metabolism | 0.028 | 0.068 |
| Propanoate metabolism | 0.030 | 0.018 |
| Cyanoamino acid metabolism† | 0.032 | 0.074 |
| PPAR signaling pathway | 0.042 | <0.001 |

[a]Pathway activity quantifies the difference between control and DBP-treated samples from Liu et al. (2005) (see Eq. 6-6). PAL is the pathway activity level for both the control and treated samples (see Eq. 6-2). The statistical significance of PA and PAL values are evaluated through a randomization procedure. The *p*-value of PAL is used as an additional filtering process to eliminate potentially nonactive pathways.
[b]The *p*-value of the PA is computed as the fraction of the randomized PA that exceeded the actual PA. In the event that the PA of the randomly generated matrices exceeds the actual PA by more than 5 % of the randomization process, then the actual PA is attributed to a random variable ($p$-value < 0.05).
[c]The *p*-value of PAL quantifies the significance of fraction of the overall gene expression captured by PAL. It is computed as the fraction of the randomized $f_p$ exceeding the actual $f_p$. In the event that the PA of the randomly generated matrices exceed the actual PA by more than 5 % of the randomization process, then the actual PA is attributed to a random variable ($p$-value < 0.05).
PA, pathway activity; PAL, pathway activity level.

To explore the biological significance of the active pathways, a metabolic pathway network of the active pathways illustrating their connections via metabolites was built (Figure 6-2). This process includes the integration of the statistical outcome of the pathway activity analysis and the relationships among these pathways by querying the KEGG database. After DBP *in utero* exposure, the pathways related to cholesterol biosynthesis exhibit more significant changes in their gene expression compared to the rest of the active pathways. This finding is consistent with the hypothesis that an early decrease in T level might be due to cholesterol unavailability (Thompson et al., 2005).

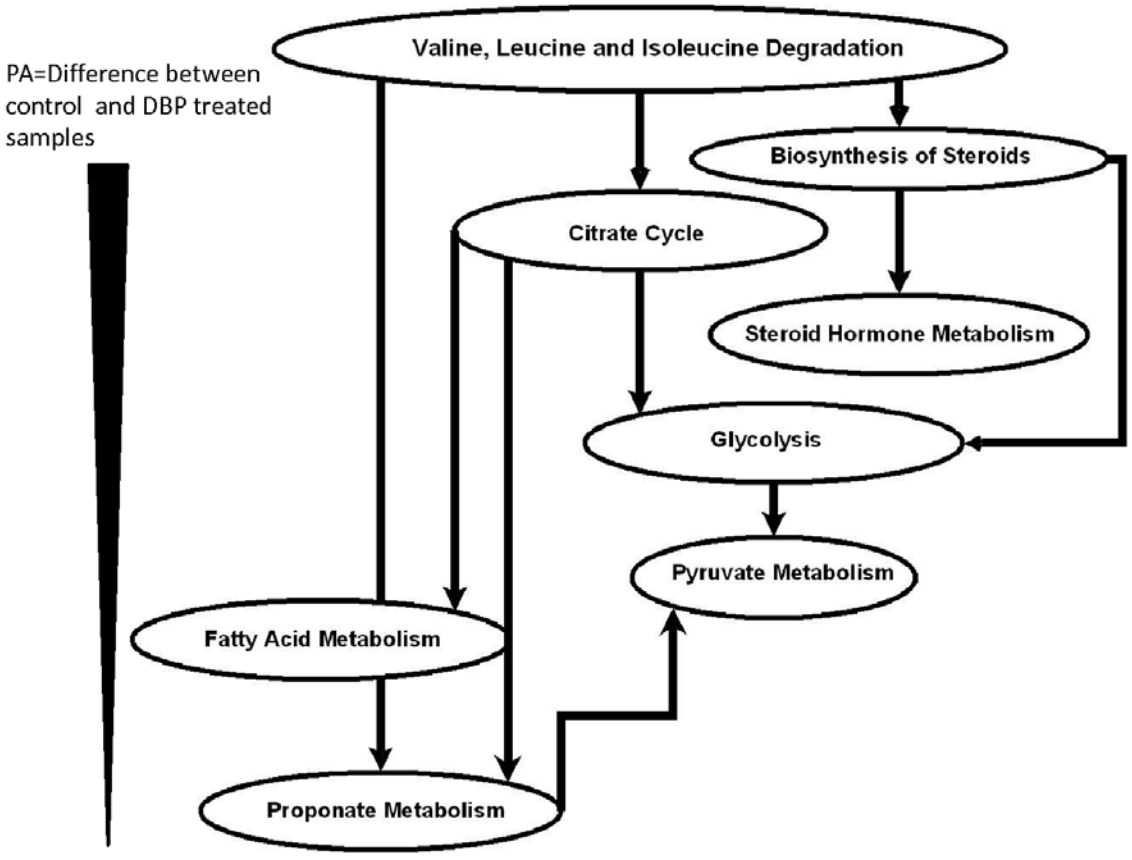


**Figure 6-2. Metabolic pathway network for DBP (Liu et al., 2005 data) using the pathway activity method and the KEGG database.** Active pathways connected to each other via metabolites are ordered from the most active pathway (top of the figure) to the less active pathways (bottom of the figure). The connections between the active pathways were retrieved from KEGG (Kanehisa and Goto, 2000).

We explored the contribution of DEGs to the pathway activity for a given pathway (Figure 6-3 A, B, C, and D). The pathway activity of each pathway is calculated by adding genes one-by-one starting with the gene with the highest SNR and adding genes sequentially in the order of their SNR until all genes in the pathway have been added. Figure 6-3 A and B illustrate examples of active pathways, whereas Figure 6-3 C and D are examples of pathways that were not identified as active in our analysis. For pathways that were identified as active or not active, the cumulative pathway activity value undergoes a decrease as genes of lower SNRs are added. Yet for the active pathways, the cumulative pathway activity remains high enough to be statistically significant. For pathways identified as not active, the cumulative pathway activity reaches a low level when all of the genes are added. Accordingly, their pathway activity value is not statistically significant. The four pathways are composed of a similar number of genes; therefore, the number of genes in the pathway is not an issue in this comparison. We hypothesize that there is a subset of genes that maintain the pathway activity value high enough within active pathways, even when all genes are added. The cumulative behavior of this subset enables us to differentiate the active and nonactive pathways. Differentially expressed genes in active pathways are defined as "informative genes" (see Table B-1). We identified a relatively small number of genes as informative, and these may represent genes that DBP has most greatly affected.

One of our goals was to utilize existing DBP genomic data to develop a gene network model useful to risk assessment. Gene network models illustrate interactions between genes and their products (e.g., mRNA, proteins). We used IPA software to construct a gene network model after DBP *in utero* exposure. IPA adds nodes (i.e., genes) to the input gene list (i.e., informative genes) and then, builds edges (i.e., relations) based on the literature. The interactions among the informative genes from the Liu et al. (2005) data were retrieved using IPA, and the resulting preliminary gene network model is shown in Figure 6-4.

### 6.2.2. Developing a Temporal Gene Network Model

The Thompson et al. (2005) study was selected to develop a temporal gene network because it was the only available time-course study. The study had the advantages of using the rat Affymetrix chip, which has ~30,000 gene transcripts represented, and availability of the data (i.e., kindly provided by Dr. Kevin Gaido). In the study, animals were exposed to DBP for 0.5,

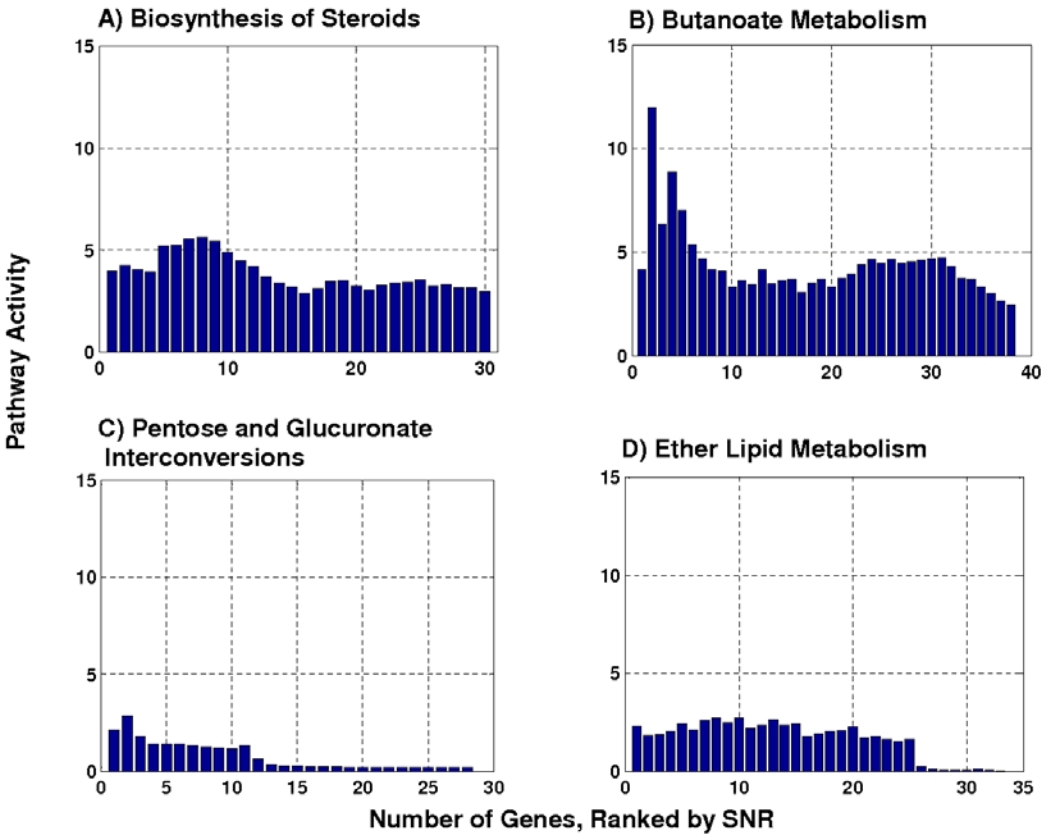


**Figure 6-3. The relationship between differential expression of individual genes and pathway activity using the Liu et al. (2005) DBP data.** The pathway activity of a given pathway is first evaluated using the gene that has the highest SNR. Subsequently, the genes are added in the order of their SNR, from highest to lowest. Pathways identified as active for DBP, such as biosynthethis of steroids (A) and butanoate metabolism (B), maintain high pathway activity values even when all genes in the pathway are added. Alternatively, pathways not identified as active for DBP such as pentose and glucuronate interconversions (C) and ether lipid metabolism (D), exhibit a decrease in pathway activity as the less discriminating genes (i.e., those with a lower SNR value) are added.

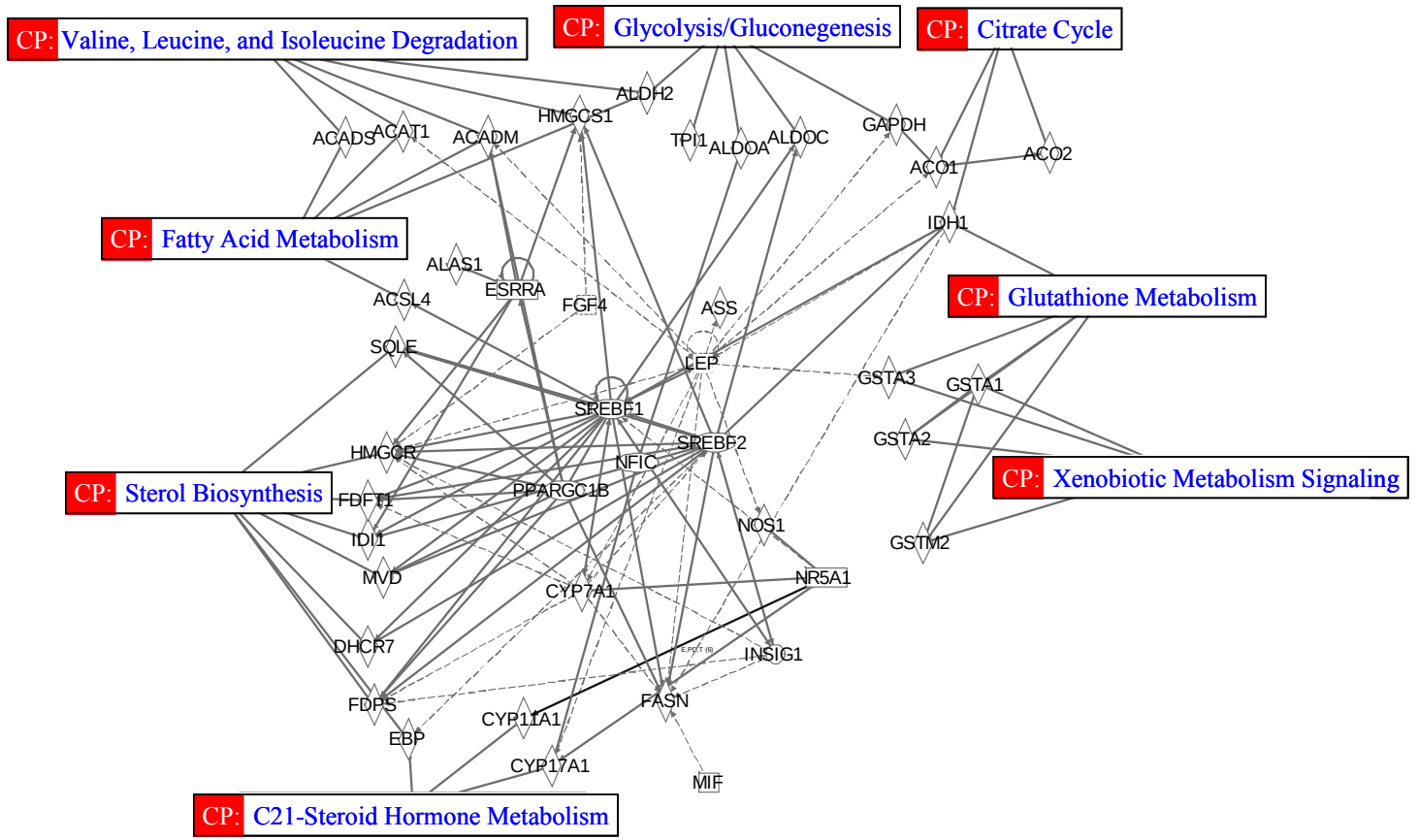


**Figure 6-4. A gene network for DBP data of Liu et al. (2005) generated using Ingenuity Pathway Analysis (IPA).** The figure illustrates the interactions among informative genes following *in utero* DBP exposure in the rat testis from Liu et al. (2005). Interactions among genes (shown in Appendix B, Table B-1) are derived from the Ingenuity database. Genes or gene products are represented as nodes. Diamonds, enzymes; Horizontal ovals, transcription regulators; Squares, cytokines; Rectangles, nuclear receptors. Solid lines represent direct relationships between nodes (i.e., molecules that make physical contact with each other, such as binding or phosphorylation). Dashed lines represent indirect interactions (i.e., not requiring physical contact between the two molecules, such as signaling events). CP, Canonical pathway.

1, 2, 3, 6, 12, 18, or 24 hours before sacrifice on GD 19. The limitations of the Thompson et al. (2005) study include (1) the dosing was initiated on GD 18, late in the critical window, and (2) the shortest duration exposure (30 minutes) began at the latest developmental time (i.e., GD 19); thus, developmental stage and duration of exposure do not coincide (see Chapter 5). Given this caveat, we utilized the available to test algorithms to build a prototype of a temporal gene network model.

We used the pathway activity level method described earlier to identify biologically active pathways at each time point. We evaluated the informative genes at each time point and the resulting preliminary temporal gene network, based on the Thompson et al. (2005) data, is shown in Figure 6-5. The analysis showed a preponderance of signaling pathways such as JAK/STAT, PPAR, and MAPK perturbed at the earlier exposure durations. After the longest DBP exposures (18 hours), the metabolic pathways, including amino acid metabolism, lipid metabolism, and carbohydrate metabolism, were affected. Thompson et al. (2005) hypothesized that the decrease in T level after a short duration of DBP exposure might be due to cholesterol unavailability and their findings support this hypothesis. To have a complete understanding of the temporal sequence of gene expression and pathway affect events after *in utero* DBP exposure, data from an exposure-duration series across the entire critical window of exposure are needed.

## **6.3.** EXPLORATORY METHODS: MEASURES OF INTERSPECIES (RAT-TO-HUMAN) DIFFERENCES IN TOXICODYNAMICS

The goal of this section is to address whether genomic and mechanistic data could inform the interspecies (rat-to-human) differences TD for one of the DBP MOAs reduced fetal testicular T (one of the DBP case-study questions). Although progress has been made in understanding the MOAs of chemical toxicants, it is important to evaluate the mechanistic relevance of these MOAs to humans. The genomic data set for DBP does not include human genomic data of any type, including studies from *in vitro* cell lines. Even if such data were available, extrapolation of *in vivo* data (rat genomic) to *in vitro* data (human genomic) may confound the ability to generate accurate interspecies comparison. In the absence of DBP genomic data in human cell lines, we considered genetic sequence data and other data from rats and humans for making species comparisons. It is significant that the role of T in male reproductive development during sexual

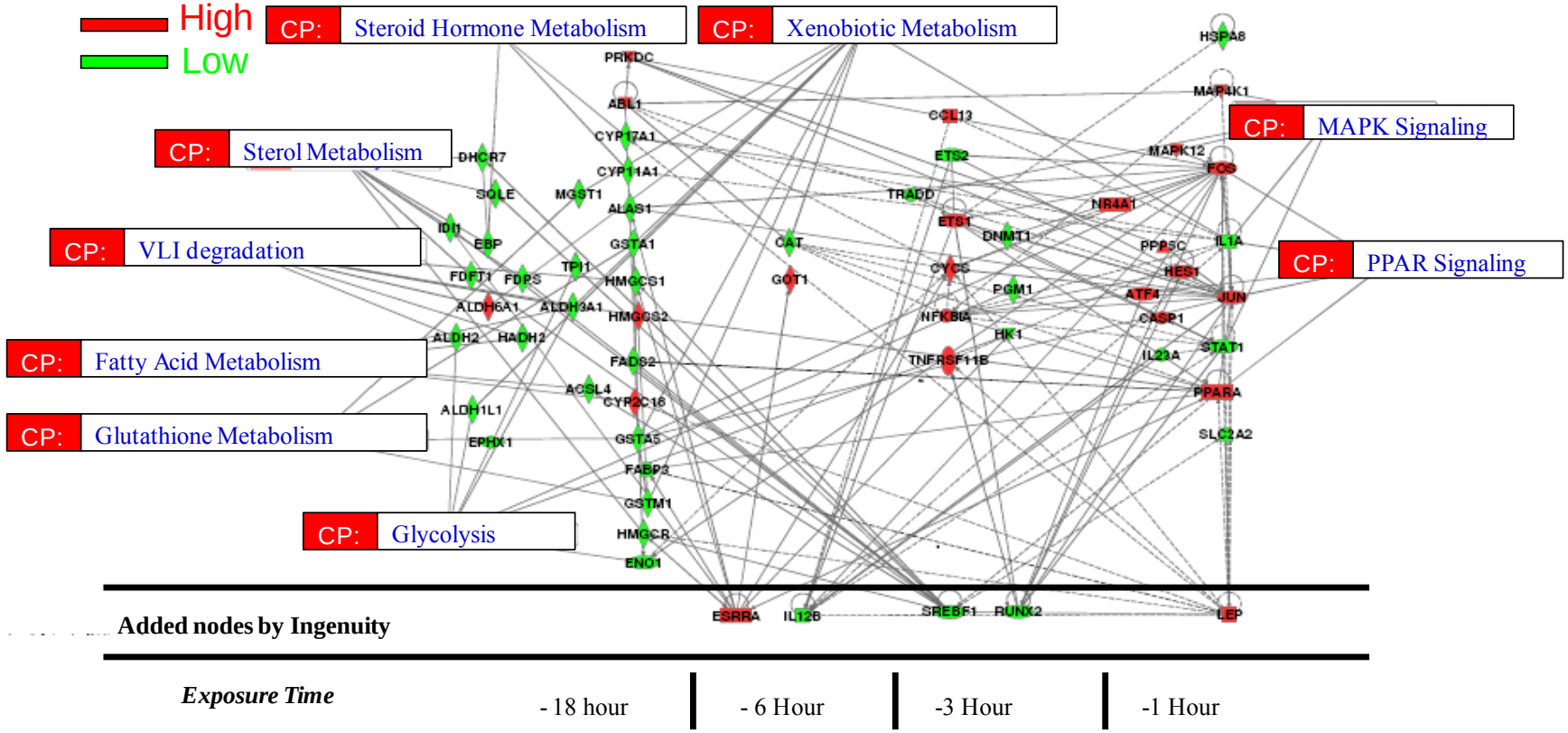


**Figure 6-5. A temporal gene network model created by IPA from the informative gene list based on time-course data after *in utero* DBP exposure (Thompson et al., 2005).** The informative genes were evaluated at each time point and mapped onto a global molecular network using the Ingenuity Pathways Knowledge Base. Diamonds, enzymes; Horizontal ovals, transcription regulators; Squares, cytokines; Rectangles, nuclear receptors. Solid lines represent direct relationships (also called edges) between nodes (i.e., molecules that make physical contact with each other, such as binding or phosphorylation). Dashed lines represent indirect interactions (i.e., not requiring physical contact between the two molecules, such as signaling events). CP, Canonical pathway. Low (green), downregulated expression with respect to control. High (red), upregulated expression with respect to control.

differentiation is conserved among vertebrates, thus providing a measure of human relevance of the reduced fetal testicular T observed in the rat after *in utero* DBP exposure.

Phylogenetic analysis, the reconstruction of evolutionary relations, is based on shared, derived characters presumed to have a common origin. Taxonomy of organisms is one method for determining species relatedness. However, since DBP perturbs the activity of the steroidogenesis pathway and leads to the decreased T MOA for DBP, we were interested in developing metrics by comparing this pathway between the rat (for which we have data) and human. Previous phylogenetic analyses of individual pathways have included assessing: the number of common enzymes and their conservation across different species (Forst, 2002; Forst and Schulten, 1999); the topology of the underlying enzyme-enzyme relational graphs including their sequence conservation (Heymans and Singh, 2003); and the intersection of compounds, reactions, and enzymes across species (Clemente et al., 2005).

We reconstructed the phylogenetic relationships for biosynthesis of steroids among eight species based on enzyme presence (Forst and Schulten, 1999; see Figure 6-6). The enzyme presence method is based on information available in the KEGG database about the presence of an enzyme (defined as catalyzing a specific reaction) in the pathway for a given species. As a result, a pathway topology can be represented and compared across species. In this representation of pathways, a vector containing binary information (where "1" is for presence, "0" is for absence of the enzyme) is created for a given pathway. Then, the similarity between pathways for two different species is defined as the ratio of the number of common enzymes to the number of unique enzymes. The results suggest that the steroidogenesis pathway is quite similar between rat and human. Further, we found that the species differences based on enzyme presence were different from those based on the NCBI taxonomy (Sayers et al., 2008) of the organisms, which is not surprising based on previous findings (Searls, 2003). In order to utilize more complete information about a pathway, cross-species pathway comparisons should include other biologically relevant information such as gene regulatory information and pathway interactions.

Sequencing of the human, mouse, and rat genomes and their comparison has increased our understanding of cross-species similarities and differences in genes and proteins. Co-expressed genes across multiple species are most likely to have a conserved function. The rat genome project reported that almost all human genes known to be associated with disease have

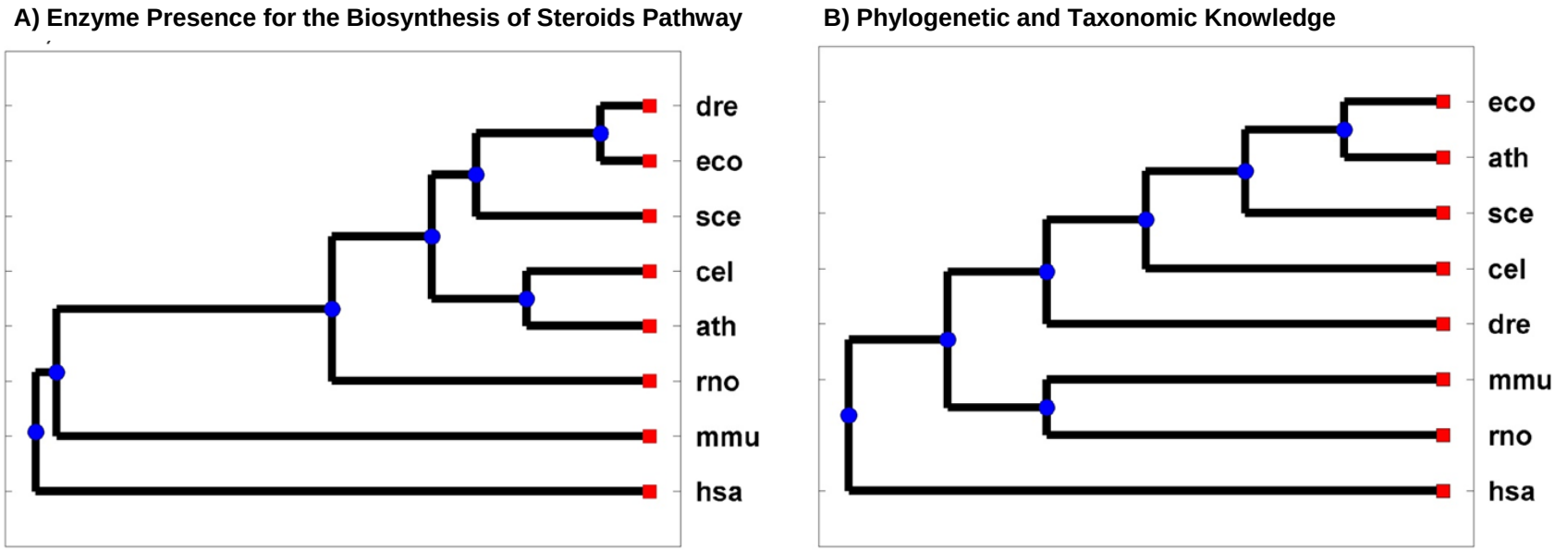


**Figure 6-6. The phylogenetic relations among eight organisms based on enzyme presence, for the biosynthesis of steroids pathway, and based on information available on the NCBI taxonomy website (Sayers et al., 2008).** Panel A shows the results of evaluating the phylogenetic relations for the biosynthesis of steroids pathway, based on enzyme presence (KEGG database), among eight model species (hsa, *Homo sapiens*; mmu, Mouse; rno, Rat; dre, Zebra fish; ath, Arabidopsis; cel, *C. elegans*, sce, Yeast; eco, *E. coli*). Panel B shows the phylogenetic relations among the same eight organisms based on taxonomic and phylogenetic information retrieved from the NCBI taxonomy database (http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/index.cgi).

orthologous genes in the rat genome, and that the human, mouse, and rat genomes are approximately 90% homologous (Gibbs et al., 2004). While the function of certain genes and their involvement in disease might not be conserved across species, the function of a pathway is likely to be more highly conserved among species that perform similar functions (Fang et al., 2005). Thus, cross-species pathway conservation metrics may be more useful.

Similarity among species can be investigated by phylogenomics analysis that involves a comparison of genes and gene products across a number of species, characterizing homologues and seeking further insights about evolutionary relationships. Analyzing the similarities between phylogenetic gene trees and their associated protein trees can reveal additional information. For example, a reconstruction of the CYP2A family of cytochrome P450 enzymes indicates that the rat liver isoform (CYP2A1) has diverged significantly from the human (CYP2A6) and mouse (CYP2A4) enzymes, having a distinct branch of the tree rooted outside the rest of the family (Searls, 2003). This considerable deviation is associated with a well-known functional shift that the rat enzyme causes the coumarin to be metabolized to a hepatotoxic epoxide, whereas the human and mouse enzymes act on the same substrate by way of a more harmless hydroxylation.

The same principles can be extended to amino acid sequence comparisons for the genes that make up a pathway. Utilizing the predicted amino acid sequence information for genes in the steroidogenesis pathway from rats and humans, we evaluated the similarity among this set of genes. Preliminary results suggest that proteins involved in the biosynthesis of steroids are highly conserved across rats and humans, with the average sequence similarity of enzymes between human and rat being ~87% as presented in Table 6-2. However, it is difficult to unequivocally determine a "high" versus "low" degree of conservation for the genes in this pathway—especially in light of the fact that events important to the effect of DBP on steroidogenesis are not well-understood. For example, initiating event after DBP exposure is not known. Additionally, there are likely differences between identifying a gene that is statistically highly conserved versus understanding whether or not the biologically meaningful regions of the predicted protein sequence, active sites, are conserved. However, endocrinological, developmental, and genetic studies in many vertebrate species indicate that the role of androgens is highly conserved across vertebrates, as androgens are critical for sexual differentiation in the male. Thus, taken together, this information suggests a high conservation of steroidogenesis and androgen synthesis in rats and humans.

**Table 6-2. The amino acid sequence similarity of the enzymes in the steroidogenesis pathway between rat and human.**

| Gene symbol | Entrez gene ID | mRNA and protein IDs | Human homolog IDs | Identities[a] | Positives[b] | Gaps[c] |
|---|---|---|---|---|---|---|
| *Dhcr7* | 64191 | NM_022389.2→NP_071784.1 | Q9UBM7 | 412/475 (86%) | 443/475 (93%) | 4/475 (0%) |
| *Idi1* | 89784 | NM_053539.1→NP_445991.1 | AF003835 | 196/227 (86%) | 215/227 (94%) | 0/227 (0%) |
| *Fdps* | 83791 | NM_031840.1→NP_114028.1 | M34477 | 301/353 (85%) | 326/353 (92%) | 0/353 (0%) |
| *Fdft1* | 29580 | NM_019238.2→NP_062111.1 | AAP36671 | 356/413 (86%) | 393/413 (95%) | 0/413 (0%) |
| *Hmgcr* | 25675 | NM_013134.2→NP_037266.2 | AAH33692 | 738/890 (82%) | 768/890 (86%) | 58/890 (6%) |
| *Mvd* | 81726 | NM_031062.1→NP_112324.1 | AAP36301 | 338/398 (84%) | 357/398 (89%) | 1/398 (0%) |
| *Sqle* | 29230 | NM_017136.1→NP_058832.1 | NP_003120 | 481/574 (83%) | 528/574 (91%) | 1/574 (0%) |
| *Ebp* | 117278 | NM_057137.1→NP_476478.1 | NP_002331 | 618/732 (84%) | 673/732 (91%) | 1/732 (0%) |
| *Lss* | 81681 | NM_031049.1→NP_112311.1 | NP_002331 | 618/732 (84%) | 673/732 (91%) | 1/732 (0%) |
| *Sc5d* | 114100 | NM_053642.2→NP_446094.1 | NP_008849 | 246/299 (82%) | 275/299 (91%) | 0/299 (0%) |
| *Mvk* | 81727 | NM_031063.1→NP_112325.1 | BAD92959 | 323/393 (82%) | 355/393 (90%) | 0/393 (0%) |
| *Cyp27b1* | 114700 | NM_053763.1→NP_446215.1 | NP_000776 | 413/508 (81%) | 453/508 (89%) | 7/508 (1%) |
| *Nqo1* | 24314 | NM_017000.2→NP_058696.2 | NP_000894 | 234/274 (85%) | 250/274 (91%) | 0/274 (0%) |
| *Vkorc1* | 309004 | NM_203335.2→NP_976080.1 | AAQ13668 | 83/94 (88%) | 88/94 (93%) | 0/94 (0%) |
| **Average similarity scores** | | | | 84% | 94.14% | |

**Table 6-2. (continued)**

[a]Identities, The number and fraction of total residues in the HSP which are identical.
[b]Positives, The number and fraction of residues for which the alignment scores have positive values.
[c]Gap, A space introduced into an alignment to compensate for insertions and deletions in one sequence relative to another. To prevent the accumulation of too many gaps in an alignment, introduction of a gap causes the deduction of a fixed amount (the gap score) from the alignment score. Extension of the gap to encompass additional nucleotides or amino acid is also penalized in the scoring of an alignment.

The HSP (high-scoring segment pair) is the fundamental unit of BLAST algorithm output. Alignment: The process of lining up two or more sequences to achieve maximal levels of identity (and conservation, in the case of amino acid sequences) for the purpose of assessing the degree of similarity and the possibility of homology.

Source: http://searchlauncher.bcm.tmc.edu/help/BLASToutput.html#anchor14684156.

The same principles can be extended from amino acid sequence comparisons to structures, pathways, and expression patterns.

## **6.4.** CONCLUSIONS

The exploratory projects presented in this chapter include efforts to develop methods for analyzing genomic data for use in risk assessment and examples of genomic data analyses available to the risk assessor with expertise in bioinformatics. These methods include pathway level analysis (including the newly described pathway activity method), gene network analysis, and tools to assess cross-species similarities in pathways. A summary for a less technical reader is presented below, grouped by the three objectives for the work.

1. *Explore the development of new methods for pathway analysis of microarray data for application to risk assessment.*

   Quality-control requirements for microarray study analysis for use in risk assessment are distinct from basic research. In traditional pathway level analysis, differentially expressed genes are first identified and then mapped to their respective pathways. Depending on the number of genes that map to a given pathway, the role of the pathway can be over- or underestimated. To overcome this problem, we used the pathway activity method. This method scores pathways based on the expression level of all genes in a given pathway.

   The pathway activity analysis identified valine, leucine, isoleucine (VL1) degradation, sterol biosynthesis, citrate cycle, and fatty acid metabolism as the most active pathways following DBP exposure. These findings support the hypothesis of Thompson et al. (2005), that an early decrease in T levels may be a result of cholesterol unavailability. However, for this approach to be useful, knowledge of tissue-specific pathways is required. For example, even though bile acid biosynthesis does not take place in the testis, a pathway related to bile acid biosynthesis was identified as statistically significant in this analysis. This method shows promise for use in risk assessment.

2. *Utilize existing DBP genomic data to develop a gene network model for use in risk assessment.*

   Determining a sequence of gene expression changes and pathway level effects over time can be very useful for understanding the temporal sequence of critical biological events perturbed after chemical exposure, and thus, useful to a risk assessment. We developed a method for developing a gene network model for DBP based on the available data. The availability of time-course data (Thompson et al., 2005) enabled our group to model the series of events that occurred between exposure to DBP and the onset of toxic reproductive outcomes. However, given the limitations of the Thompson et al. (2005) study design, we did not draw conclusions about genes and pathways affected over time

for DBP. Instead, the Thompson et al. (2005) data was used to build a prototype of a temporal gene network model and thus, the exercise allowed us to develop methods for analyzing time-course data.

3. *Utilize genomic and other molecular data to address the Case Study Question: Do the toxicogenomic data inform interspecies differences in TD?*

   Extrapolation from animal-to-human data is critical for establishing human relevance of MOA(s) in risk assessment. Co-expressed genes across multiple species could have a conserved function. The human, mouse, and rat genomes have been reported to be 90% homologous (Gibbs et al., 2004). However, because it is not certain whether the function of a specific gene is conserved across species, conservation of pathways across species can be one important factor in establishing cross species concordance of one or more MOAs. In addition, a common critical role of androgens in both rodent and human male development of reproductive organs has been well-established.

   Using the available DNA, sequence, and protein similarity data for the steroidogenesis pathway, we used three different methods to assess rat-to-human conservation as metrics that may inform the interspecies differences in TD for one MOA, the reduced fetal testicular T. The pathways for the biosynthesis of steroids have similarity between human and rat. Comparing the predicted amino acid sequences for the steroidogenesis pathway genes, we found that the average sequence similarity between rat and human is ~87%, and the average promoter region similarity of genes is 52%. Some of the challenges in using similarity scores to estimate the cross-species relevance of a MOA are described (see Section 6.3).

In summary, the preliminary analytical efforts described in this chapter address and raise a number of issues about the best approaches for analyzing microarray and other genomic data for risk assessment purposes. Traditional pathway analysis methods, while useful, also restrict the incorporation of all genes for determining relevant pathways that are affected by DBP. There is a substantial amount of background noise generated in a typical microarray experiment (i.e., gene expression variability even among the controls; see Smith, 2001). For use in risk assessment, it is important to be able to identify and separate the signal from the noise. Innovative approaches, such as the pathway activity method described in this chapter, may provide more confidence when evaluating microarray data for use in risk assessment. These efforts reveal some of the promises and challenges of analyzing and interpreting genomic data for application to risk assessment.